Dear Sir or Madam, May I introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer (Supplementary Material)

Sudha Rao
University of Maryland, College Park
raosudha@cs.umd.edu

Joel Tetreault
Grammarly
joel.tetreault@grammarly.com

1 Introduction

In this supplementary material, we add additional details supporting the dataset §2, models §3 and results §4 we introduce in the main paper. In §4, we also discuss results of our models on the *formal* to *informal* task.

2 Dataset

In Section 3.2 of the main paper, we perform a qualitative analysis to understand the types of edits people made for making a sentence more formal. Table 1 shows the frequency of each types of edits and example sentence pairs for the same. In addition, for a subset of the categories for which we can count the edits automatically, we show the frequency of edits on the entire train split of our GYAFC dataset where we observe a higher percentage of capitalization and punctuation edits as compared to manual counting and a much higher percentage of normalizations.

3 Models

3.1 Rule-based Model

We use the analysis described in Section 3.2 of the main paper to construct the following set of rules to automatically make an informal sentence more formal:

Capitalization: We capitalize the first letter of a sentence, we capitalize the pronoun 'I' and we capitalize proper nouns by identifying words with parts of speech NNP or NNPS.

Lowercase words with all upper cases: In several informal sentences, words are often capitalized for emphasis, e.g. *"ARE YOU KIDDING ME????"* We lowercase such sentences or words.

Expand contractions: Informal sentences contain contractions like *'wasn't', 'haven't',* etc. We handcraft a list of expansions for all such contractions.

Replace slang words: Informal sentences contain slang word usage like *'juz', 'wanna',* etc. We handcraft a list of slang replacements.

Replace swear words: Informal sentences frequently contain swear words. We handcraft a list of swear words and replace all but their first character with asterisks. Example, *'suck'* is replaced with *'s***'*.

Remove character repetition: Informal sentences contain several instances of repeated characters for emphasis. For example, *'nooooo', 'yayyyyy', '!!!!', '???'*. We use regular expressions to replace such repeated occurrences with a single occurrence.

The rule-based model for the second direction of style transfer i.e. from *formal* to *informal* consists of the same rules as above but in the reverse direction. The rules of capitalization, contractions and slang usage are applied always whereas the rules of uppercasing and character repetition are applied proportionally to their occurrences in the GYAFC train split.

3.2 PBMT Model

The different ideas combined together to obtain the PBMT model in Section 4.2 of the main paper are described in detail below:

PBMT on rule-based: When we use the ∼50K sentence pairs in our GYAFC train set to train a baseline PBMT model, we observe that the model mainly learns to replicate the rules we crafted in our rule-based approach. Hence, in order to force the PBMT model to learn generalizations beyond the rules, we train the PBMT model on the output of our rule-based approach such that the source side of the parallel data is now the output of the rule-based approach. For all of our subsequent models we use this parallel data.

Category	Manual	Auto	Original Informal	Formal Rewrite
Paraphrase	47%	–	he iss **wayyyy hottt**	He is **very attractive**.
Capitalization	46%	51%	**yes**, exept for episode **iv**.	**Yes**, but not for episode **IV**.
Punctuation	40%	69%	I've watched it and it is AWESOME**!!!!**	I viewed it and I believe it is a quality program**.**
Delete fillers	26%	–	**Well...** Do you talk to that someone much?	do you talk to that person often?
Completion	15%	–	**Haven't seen** the tv series, but R.O.D.	**I have not seen** the television series, however I have seen the R.O.D
Spelling	14%	–	that page did not give me **viroses** (i think)	I don't think that page gave me **viruses**.
Contractions	12%	8%	I **didn't** know they had an HBO in the 80's	I **did not** know HBO existed in the 1980s.
Normalization	10%	61%	my exams **r** not over yet	My exams **are** not over yet.
Lowercase	7%	8%	But you will **DEFINALTELY** know when you are in love!	You will **definitely** know when you are in love.
Split Sentences	4%	–	it wouldnt be a word, it would be me singing operah.	It would not be a word. It would be a singing opera.
Repetitions	2%	5%	i'd find out what **reallllllllllllly** happened to marilyn monroe	I would determine what **really** happened to Marilyn Monroe.

Table 1: Categories of frequent edits calculated using manual and automatic counting and examples of each. Note that the categories are not mutually exclusive.

Train	Test	$Spearman(\rho)$
PT16	PT16	0.68
PT16	E&M	0.39
PT16	F&R	0.38
E&M	E&M	0.56
F&R	F&R	0.51

Table 2: Spearman rank correlation between formality classifier predictions and human judgments on 10-fold cross validation on 5000 sentences.

Self-training: The amount of parallel data in our *formality* dataset is orders of magnitude smaller than the amount typically used for training translation models. We therefore increase the size of our train set by way of *self-training* where we use the PBMT model to translate the large number of in-domain sentences from GYAFC belonging to the the source style and use the resultant output to retrain the PBMT model.

Sub-selection using Edit Distance: A large portion of the training data obtained via *self-training* consists of parallel sentences where the two sides are almost identical. In order to push the PBMT model towards translations that involve higher number of edits, we sub-select the additional training data generated using *self-training* to include only those where the edit distance between the two sides is more than 10. Further, to ensure the equal proportion of the original parallel data and the additional data, we up-weight the original parallel data via duplication.

Larger Language Model with Data Selection: The Yahoo Answers corpus contains a large number of target style sentences spanning across different domains that we could potentially use to train a larger language model, but at the cost of domain mismatch. To sub sample from large out-of-domain data, we use intelligent data selection method and train a language model on sentences that are closer to the target style in-domain data.

Model	Training data
	E&M
PBMT Baseline	50K
PBMT Combined	50K*6+0.3M
NMT Combined	50K*6+0.3M+0.7M
	F&R
PBMT Baseline	50K
PBMT Combined	50K*10+0.3M
NMT Combined	50K*10+0.6M+1.8M

Table 3: Sizes of the training data used in the different models for the two domains.

Table 3 contains the approximate sizes of the training data used in the main models across the two domains. Under the "Combined" models, the first part is duplication of the GYAFC train split, the second part is additional data obtained via self-training with sub-selection and the third part is the additional data obtained via back-translation for the "NMT Combined model".

3.3 Details of NMT model

We use the OpenNMT-py (Klein et al., 2017) toolkit with default parameters with a vocabulary size of 50K and embeddings of size 300. At test time, we replace unknown tokens with the source token that has the highest attention weight. The input word embeddings are pretrained on Yahoo Answers using GloVE (Pennington et al., 2014).

E&M	F&R
(2.33) PBMT Combined	(2.24) Rule-based
(2.40) NMT Combined	(2.28) PBMT Combined
(2.44) Rule-based	(2.36) NMT Combined
(2.52) Reference	(2.44) Reference
(2.69*) NMT Copy	(2.50*) NMT Baseline
(2.72*) NMT Baseline	(2.53*) NMT Copy

Table 4: Ranking of different models on *formal* to *informal* task. Rankings marked with * are significantly different from the rule-based ranking with $p < 0.001$.

Automatic	**Human**	**E&M**	**F&R**
BLEU	*Overall*	-0.10*	-0.01
TERp	*Overall*	0.06	0.02
PINC	*Overall*	0.17*	0.15*
Formality	*Formality*	0.55*	0.57*
Fluency	*Fluency*	0.46*	0.43*
Meaning	*Meaning*	0.28*	0.28*
Combined	*Combined*	0.38*	0.38*

Table 5: Spearman rank correlation between automatic metrics and human judgments. The first three metrics are correlated with the *overall ranking* human judgments and the last four are correlated with the human judgments on the respective three axes. Correlations marked with * are statistically significant with $p < 0.001$

4 Results

4.1 Results on F&R domain

In the main paper, we include results for only the E&M domain. Here we discuss the results on the F&R domain. Table 6, similar to Table 4 in the main paper, shows the results of the models on 500 test sentences evaluated using both human judgments and automatic metrics. The main observations regarding model performances across all metrics are similar to the E&M domain. One difference is that the formality score of the original informal sentences in F&R are higher than in E&M and consequently the formality scores of the formal rewrites from both human references and model outputs are higher than in E&M.

4.2 Results on *Formal* to *Informal* task

In the main paper, we focus on the *informal* to *formal* direction of style transfer. In this section we discuss the results of our models on the other direction. It should be noted that this experimentation is fundamentally different from the first direction in a way that instead of identifying *formal* sentences from Yahoo Answers and collecting their *informal* rewrites, we reuse the data created for the first direction.

In Table 7, we show the results of the five main models on the *formal* to *informal* task. The main observation is that, in contrast to the first direction, the rule-based model beats all other models across all three criteria of *formality*, *fluency* and *meaning* as per human judgments and automatic metrics (with the exception of meaning automatic metric where PBMT Combined beats rule-based). NMT Combined and PBMT Combined win as per BLEU and TERp and NMT Baseline wins as per PINC. As in Section 6.3 of the main paper, in Table 5, we report the correlation of these metrics with human judgments. In contrast to the first direction, the formality classifier obtains a higher correlation which might be because the classifier is trained on informal data and so it is better at assessing informal model outputs than formal model outputs. The fluency and meaning correlation are about the same. In contrast, BLEU, TERp and PINC all three correlate very poorly with the overall ranking. This difference might be explained by the fact that informal reference rewrites vary highly in that there are much higher number of ways of making a sentence more informal as compared to making it more formal. Therefore, metrics that make use of references might be ill-suited for this style transfer task.

In Table 8, we show some sample model outputs for the E&M and F&M domain sentences. We can see that rule-based method uses simple lexical transformations like 'just' to 'juz', 'you' to 'u', 'because' to 'cuz', 'love' to 'luv', etc and wins over other models. The intent of evaluating our models on this direction of task was to understand how well the same model would do the reverse task. We find that the second direction has different set of challenges and requires models that cater to those specifically if we wish to beat simple rule-based methods.

References

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, pages 67–72. https://doi.org/10.18653/v1/P17-4012.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

Model	Formality		Fluency		Meaning		Combined		Overall		
	Human	PT16	Human	H14	Human	HE15	Human	Auto	BLEU	TERp	PINC
Original Informal	*-0.90*	*-0.80*	*3.92*	*3.09*	–	–	–	–	52.61	0.34	0.00
Formal Reference	0.41	0.22	4.43	3.74	4.56	3.54	5.68	4.76	100.0	0.37	67.83
Rule-based	-0.18	-0.15	4.08	3.26	**4.82**	**4.29**	5.41	4.78	68.17	0.27	26.89
PBMT Combined	0.05*	0.11*	4.15	3.48*	4.65*	4.02*	5.45	4.88*	**74.32***	**0.25***	44.77*
NMT Baseline	0.19*	0.13*	4.18	**3.56***	3.88*	3.91*	5.20*	**4.89***	69.09*	0.31*	**51.00***
NMT Copy	**0.33***	**0.15***	4.21*	3.55*	3.97*	3.88*	5.30	4.88*	69.41	0.30	50.93
NMT Combined	0.20*	0.10*	**4.27***	3.45*	4.69*	4.06*	**5.57***	4.88*	**74.60***	**0.24***	41.52*

Table 6: Results of models on 500 test sentences for *informal* to *formal* task evaluated using human judgments and automatic metrics for three criteria of evaluation: formality, fluency and meaning preservation on the F&R domain. Scores marked with * are significantly different from the rule-based scores with $p < 0.001$.

System	Formality		Fluency		Meaning		Combined		Overall		
	Human	PT16	Human	H14	Human	HE15	Human	Auto	BLEU	TERp	PINC
Entertainment & Music											
Original Formal	*0.73*	*0.26*	*4.50*	*3.39*	–	–	–	–	*29.54*	*0.83*	*0.00*
Informal Ref	-0.88	-0.89	4.00	3.00	3.92	3.09	0.97	0.80	100.0	0.64	81.14
Rule-based	**-0.96**	**-0.88**	**3.92**	**2.99**	**4.80**	3.95	**0.55**	**0.46**	22.41	0.85	47.00
PBMT Combined	-0.16*	-0.15*	4.15*	3.32*	4.56*	**4.28***	1.03*	0.73*	32.66*	0.80*	37.38*
NMT Baseline	-0.59*	-0.61*	3.94	3.32*	3.55*	3.90	1.19*	0.73*	29.87*	0.83*	**58.38***
NMT Copy	-0.53*	-0.53*	3.98	3.30*	3.53*	3.96	1.24*	0.72*	31.29*	0.80*	55.42*
NMT Combined	-0.34*	-0.39*	4.08*	3.27*	4.32*	4.12*	1.03*	0.69*	**34.07***	**0.78***	42.31
Family & Relationships											
Original Formal	*0.73*	*0.36*	*4.47*	*3.49*	–	–	–	–	*28.64*	*0.84*	*0.00*
Informal Ref	-0.60	-0.87	4.09	3.21	3.82	3.25	1.15	0.85	100.0	0.65	80.69
Rule-based	**-1.05**	**-0.94**	**3.89**	**2.93**	**4.46**	3.98	**0.65**	**0.40**	19.56	0.87	45.95
PBMT Combined	-0.17*	-0.19*	4.18*	3.36*	4.45	**4.43***	1.09*	0.68*	31.20*	0.82*	26.74*
NMT Baseline	-0.16*	-0.46*	4.08*	3.34*	3.80*	4.11	1.30*	0.71*	**34.69***	**0.78***	**48.68**
NMT Copy	-0.19*	-0.36*	4.08*	3.34*	3.74*	4.14	1.32*	0.73*	33.76*	**0.78***	47.52
NMT Combined	-0.12*	-0.34*	4.17*	3.33*	4.22*	4.26*	1.19*	0.69*	33.57*	0.79*	36.76

Table 7: Results of models on 500 test sentences for *formal* to *informal* task evaluated using human judgments and automatic metrics for three criteria of evaluation: formality, fluency and meaning preservation on E&M and F&R domains. Scores marked with * are significantly different from the rule-based scores with $p < 0.001$.

Entertainment & Music	
Original Formal	I am just glad they didn ’t show us the toilets .
Reference Informal	IM GLAD THEY PASSED THE TOILETS
Rule-based	i am juz glad they didn ’t show us the toilets .....
PBMT Combined	I am just glad they didnt show us the toilets .
NMT Baseline	I ’m just glad they didn ’t show us the restroom .
NMT Copy	I ’m just glad they didn ’t show us the brids .
NMT Combined	I ’m just glad they didn ’t show us the toilets .
Family & Relationship	
Original Formal	Hopefully , you married your husband because you love him .
Reference Informal	you married your hubby hopefully because you love him .
Rule-based	hopefully , u MARRIED ur husband coz u luv him .....
PBMT Combined	hopefully , you married your husband because you love him .
NMT Baseline	you married your husband because you love him .
NMT Copy	Hopefully you married your husband because you love him .
NMT Combined	Hopefully you married your husband because you love him .

Table 8: Sample model outputs with references from both E&M and F&R domains on the *formal* to *informal* task